

Research Article

Microphone Diversity Combining for In-Car Applications

**Jürgen Freudenberger, Sebastian Stenzel (EURASIP Member),
and Benjamin Venditti (EURASIP Member)**

*Department of Computer Science, University of Applied Sciences Konstanz, Hochschule Konstanz, Brauneggerstr. 55,
78462 Konstanz, Germany*

Correspondence should be addressed to Jürgen Freudenberger, juergen.freudenberger@htwg-konstanz.de

Received 1 August 2009; Revised 23 January 2010; Accepted 17 March 2010

Academic Editor: Ivan Tashev

Copyright © 2010 Jürgen Freudenberger et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper proposes a frequency domain diversity approach for two or more microphone signals, for example, for in-car applications. The microphones should be positioned separately to insure diverse signal conditions and incoherent recording of noise. This enables a better compromise for the microphone position with respect to different speaker sizes and noise sources. This work proposes a two-stage approach. In the first stage, the microphone signals are weighted with respect to their signal-to-noise ratio and then summed similar to maximum ratio combining. The combined signal is then used as a reference for a frequency domain least-mean-squares (LMS) filter for each input signal. The output SNR is significantly improved compared to coherence-based noise reduction systems, even if one microphone is heavily corrupted by noise.

1. Introduction

With in-car speech applications like hands-free car kits and speech recognition systems, speech is corrupted by engine noise and other noise sources like airflow from electric fans or car windows. For safety and comfort reasons, hands-free telephone systems should provide the same quality of speech as conventional fixed telephones. In practice however, the speech quality of a hands-free car kit heavily depends on the particular position of the microphone. Speech has to be picked up as directly as possible to reduce reverberation and to provide a sufficient signal-to-noise ratio. The important question, where to place the microphone inside the car, is, however, difficult to answer. The position is apparently a compromise for different speaker sizes, because the distance between microphone and speaker depends significantly on the position of the driver and therefore on the size of the driver. Furthermore, noise sources like airflow from electric fans or car windows have to be considered. Placing two or more microphones in different positions enables a better compromise with respect to different speaker sizes and yields more noise robustness.

Today, noise reduction in hands-free car kits and in-car speech recognition systems is usually based on single

channel noise reduction or beamformer arrays [1–3]. Good noise robustness of single microphone systems requires the use of single channel noise suppression techniques, most of them derived from spectral subtraction [4]. Such noise reduction algorithms improve the signal-to-noise ratio, but they usually introduce undesired speech distortion. Microphone arrays can improve the performance compared to single microphone systems. Nevertheless, the signal quality does still depend on the speaker position. Moreover, the microphones are located in close proximity. Therefore, microphone arrays are often vulnerable to airflow that might disturb all microphone signals.

Alternatively, multimicrophone setups have been proposed that combine the processed signals of two or more separate microphones. The microphones are positioned separately (e.g., 40 to 80 cm apart) in order to ensure incoherent recording of noise [5–11]. Similar multichannel signal processing systems have been suggested to reduce signal distortion due to reverberation [12, 13]. Basically, all these approaches exploit the fact that speech components in the microphone signals are strongly correlated while the noise components are only weakly correlated if the distance between the microphones is sufficiently large.

The question at hand with distributed arrays is how to combine these microphone signals with possibly rather different signal conditions? In this paper, we consider a diversity technique that combines the processed signals of several separate microphones. The basic idea of our approach is to apply maximum-ratio-combining (MRC) to speech signals, where we propose a frequency domain diversity approach for two or more microphone signals. MRC maximizes the signal-to-noise ratio in the combined signal.

A major issue for the application of maximum-ratio-combining for multimicrophone setups is the estimation of the acoustic transfer functions. In telecommunications, the signal attenuation as well as the phase shift for each transmission path are usually measured to apply MRC. With speech applications we have no means to directly measure the acoustic transfer functions. There exists several blind approaches to estimate the acoustic transfer functions (see e.g., [14–16]) which were successfully applied to dereverberation. However, the proposed estimation methods are computationally demanding.

In this paper, we show that maximum-ratio-combining can be achieved without explicit knowledge of the acoustic transfer functions. Proper signal weighting can be achieved based on an estimate of the input signal-to-noise ratio. We propose a two stage processing of the microphone signals. In the first stage, the microphone signals are weighted with respect to their input signal-to-noise ratio. These weights guarantee maximum-ratio-combining of the signals with respect to the signal magnitudes. To ensure cophasal addition of the weighted signals, we use the combined signal as reference signal for frequency domain LMS filters in the second stage. These filters adjust the phases of the microphone signals to guarantee coherent signal combining.

The proposed concept is similar to the single channel noise reduction system presented by Mukherjee and Gwee [17]. This system uses spectral subtraction to obtain a crude estimate of the speech signal. This estimate is then used as the reference signal of a single LMS filter. In this paper, we generalize this concept to multimicrophone systems, where our aim is not only noise reduction, but also dereverberation of the microphone signals.

The paper is organized as follows: In Section 2, we present some measurement results obtained in a car environment. This results motivate the proposed diversity approach. In Section 3, we present a signal combiner that achieves MRC weighting based on the knowledge of the input signal-to-noise ratios. Coherence based signal combining is discussed in Section 4. In the subsequent section, we consider implementation issues. In particular, we present an estimator for the required input signal-to-noise ratios. Finally, in Section 6, we present some simulation results for different real world noise situations.

2. Measurement Results

The basic idea of our spectral combining approach is to apply MRC to speech signals. To motivate this approach, we first discuss some measurement results obtained in a car

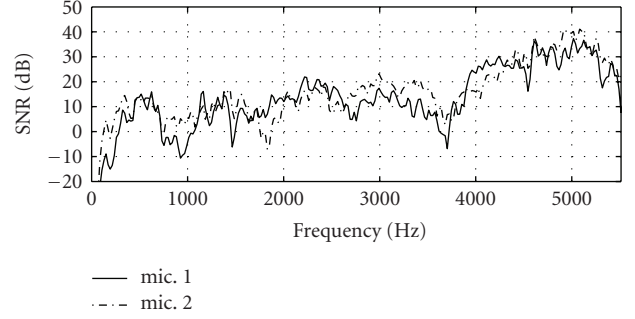


FIGURE 1: Input SNR values for a driving situation at a car speed of 100 km/h.

environment. For these measurements, we used two cardioid microphones with positions suited for car integration. One microphone (denoted by *mic. 1*) was installed close to the inside mirror. The second microphone (*mic. 2*) was mounted at the A-pillar.

Figure 1 depicts the SNR versus frequency for a driving situation at a car speed of 100 km/h. From this figure, we observe that the SNR values are quite distinct for these two microphone positions with differences of up to 10 dB depending on the particular frequency. We also note that the better microphone position is not obvious in this case, because the SNR curves cross several times.

Theoretically, a MRC combining of the two input signals would result in an output SNR equal to the sum of the input SNR values. With two inputs, MRC achieves a maximum gain of 3 dB for equal input SNR values. In case of the input SNR values being rather different, the sum is dominated by the maximum value. Hence, for the curves in Figure 1 the output SNR would essentially be the envelope of the two curves.

Next we consider the coherence for the noise and speech signals. The corresponding results are depicted in Figure 2. The figure presents measurements for two microphones installed close to the inside mirror in an end-fire beamformer constellation with a microphone distance of 7 cm. The lower figure contains the results for the microphone positions *mic. 1* and *mic. 2* (distance of 65 cm). From these results, we observe that the noise coherence closely follows the theoretical coherence function (dotted line in Figure 2) in an ideal diffuse sound field [18]. Separating the microphones significantly reduces the noise coherence for low frequencies. On the other hand, both microphone constellations have similar speech coherence. We note that the speech coherence is not ideal, as it has steep dips. The corresponding frequencies will probably be attenuated by a signal combiner that is solely based on coherence.

3. Spectral Combining

In this section, we present the basic system concept. To simplify the discussion, we assume that all signals are stationary and that the acoustic system is linear and time-invariant.

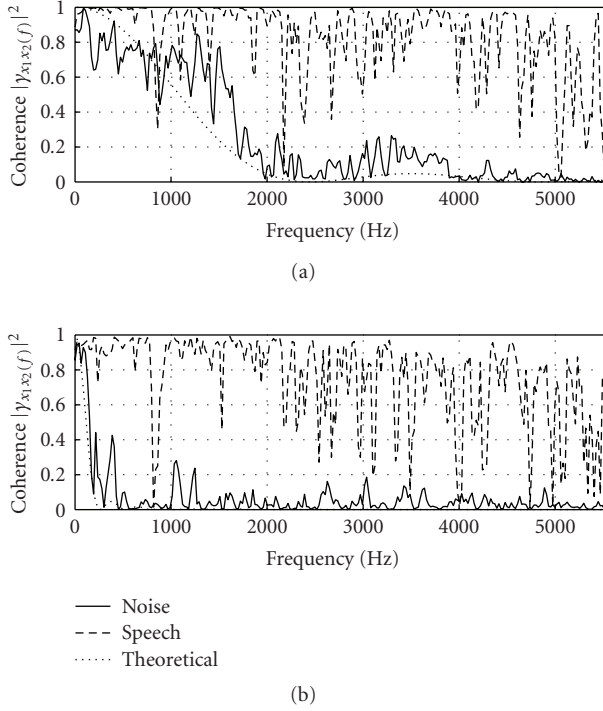


FIGURE 2: Coherence for noise and speech signals for tow different microphone positions.

In the subsequent section we consider the modifications for nonstationary signals and time variant systems.

We consider a scenario with M microphones. The microphone signals $y_i(k)$ can be modeled by the convolution of the speech signal $x(k)$ with the impulse response $h_i(k)$ of the acoustic system plus additive noise $n_i(k)$. Hence the M microphone signals $y_i(k)$ can be expressed as

$$y_i(k) = h_i(k) * x(k) + n_i(k), \quad (1)$$

where $*$ denotes the convolution.

To apply the diversity technique, it is convenient to consider the signals in the frequency domain. Let $X(f)$ be the spectrum of the speech signal $x(k)$ and $Y_i(f)$ be the spectrum of the i th microphone signal $y_i(k)$. The speech signal is linearly distorted by the acoustic transfer function $H_i(f)$ and corrupted by the noise term $N_i(f)$. Hence, the signal observed at the i th microphone has the spectrum

$$Y_i(f) = X(f)H_i(f) + N_i(f). \quad (2)$$

In the following, we assume that the speech signal and the channel coefficients are uncorrelated. We assume a complex Gaussian distribution of the noise terms $N_i(f)$. Moreover, we presume that the noise power spectral density $\lambda_N(f) = \mathbb{E}\{|N_i(f)|^2\}$ is the same for all microphones. This assumption is reasonable for a diffuse sound field.

Our aim is to linearly combine the M microphone signals $Y_i(f)$ so that the signal-to-noise ratio in the combined signal

$\hat{X}(f)$ is maximized. In the frequency domain, the signal combining can be expressed as

$$\hat{X}(f) = \sum_{i=1}^M G_i(f) Y_i(f), \quad (3)$$

where $G_i(f)$ is the weight of the i th microphone signal. With (2) we have

$$\hat{X}(f) = X(f) \sum_{i=1}^M G_i(f) H_i(f) + \sum_{i=1}^M G_i(f) N_i(f), \quad (4)$$

where the first sum represents the speech component and the second sum represents the noise component of the combined signal. Hence, the overall signal-to-noise ratio of the combined signal is

$$\gamma(f) = \frac{\mathbb{E}\left\{\left|X(f) \sum_{i=1}^M G_i(f) H_i(f)\right|^2\right\}}{\mathbb{E}\left\{\left|\sum_{i=1}^M G_i(f) N_i(f)\right|^2\right\}}. \quad (5)$$

3.1. Maximum-Ratio-Combining. The optimal combining strategy that maximizes the signal-to-noise ratio in the combined signal $\hat{X}(f)$ is usually called maximal-ratio-combining (MRC) [19]. In this section, we briefly outline the derivation of the MRC weights for completeness. Furthermore, some of the properties of maximal ratio combining are discussed.

Let $\lambda_X(f) = \mathbb{E}\{|X(f)|^2\}$ be the speech power spectral density. Assuming that the noise power $\lambda_N(f)$ is the same for all microphones and that the noise at the different microphones is uncorrelated, we have

$$\gamma(f) = \frac{\lambda_X(f) \left| \sum_{i=1}^M G_i(f) H_i(f) \right|^2}{\lambda_N(f) \sum_{i=1}^M |G_i(f)|^2}. \quad (6)$$

We consider now the term $\left| \sum_{i=1}^M G_i(f) H_i(f) \right|^2$ in the denominator of (6). Using the Cauchy-Schwarz inequality we have

$$\left| \sum_{i=1}^M G_i(f) H_i(f) \right|^2 \leq \sum_{i=1}^M |G_i(f)|^2 \sum_{i=1}^M |H_i(f)|^2 \quad (7)$$

with equality if $G_i(f) = c H_i^*(f)$, where H_i^* is the complex conjugate of the channel coefficient H_i . Here c is a real-valued constant common to all weights $G_i(f)$. Thus, for the signal-to-noise ratio we obtain

$$\begin{aligned} \gamma(f) &\leq \frac{\lambda_X(f) \sum_{i=1}^M |G_i(f)|^2 \sum_{i=1}^M |H_i(f)|^2}{\lambda_N(f) \sum_{i=1}^M |G_i(f)|^2} \\ &= \frac{\lambda_X(f)}{\lambda_N(f)} \sum_{i=1}^M |H_i(f)|^2. \end{aligned} \quad (8)$$

With the weights $G_i(f) = c H_i^*(f)$, we obtain the maximum signal-to-noise ratio of the combined signal as the sum of the signal-to-noise ratios of the M received signals

$$\gamma(f) = \sum_{i=1}^M \gamma_i(f), \quad (9)$$

where

$$\gamma_i(f) = \frac{\lambda_X(f) |H_i(f)|^2}{\lambda_N(f)} \quad (10)$$

is the input signal-to-noise ratio of the i th microphone. It is appropriate to choose c as

$$c_{\text{MRC}}(f) = \frac{1}{\sum_{j=1}^M |H_j(f)|^2}. \quad (11)$$

This leads to the MRC weights

$$G_{\text{MRC}}^{(i)}(f) = c_{\text{MRC}}(f) H_i^*(f) = \frac{H_i^*(f)}{\sum_{j=1}^M |H_j(f)|^2}, \quad (12)$$

and the estimated (equalized) speech spectrum

$$\begin{aligned} \hat{X} &= G_{\text{MRC}}^{(1)} Y_1 + G_{\text{MRC}}^{(2)} Y_2 + G_{\text{MRC}}^{(3)} Y_3 \dots \\ \hat{X} &= \frac{H_1^*}{\sum_{i=1}^M |H_i|^2} Y_1 + \frac{H_2^*}{\sum_{i=1}^M |H_i|^2} Y_2 + \dots \\ &= \frac{H_1^* (H_1 X + N_1)}{\sum_{i=1}^M |H_i|^2} + \frac{H_2^* (H_2 X + N_2)}{\sum_{i=1}^M |H_i|^2} + \dots \\ &= X + \frac{H_1^*}{\sum_{i=1}^M |H_i|^2} N_1 + \frac{H_2^*}{\sum_{i=1}^M |H_i|^2} N_2 + \dots \\ &= X + G_{\text{MRC}}^{(1)} N_1 + G_{\text{MRC}}^{(2)} N_2 + \dots, \end{aligned} \quad (13)$$

where we have omitted the dependency on f . The estimated speech spectrum $\hat{X}(f)$ is therefore equal to the actual speech spectrum $X(f)$ plus some weighted noise term.

The filter defined in (12) was previously applied to speech dereverberation by Gannot and Moonen in [14], because it ideally equalizes the microphone signals if a sufficiently accurate estimate of the acoustic transfer functions is available. The problem at hand with maximum-ratio-combining is that it is rather difficult and computationally complex to explicitly estimate the acoustic transfer characteristic $H_i(f)$ for our microphone system.

In the next section, we show that MRC combining can be achieved without explicit knowledge of the acoustic channels. The weights for the different microphones can be calculated based on an estimate of the signal-to-noise ratio for each microphone. The proposed filter achieves a signal-to-noise ratio according to (9), but does not guarantee perfect equalization.

3.2. Diversity Combining for Speech Signals. We consider the weights

$$G_{\text{SC}}^{(i)} = \sqrt{\frac{\gamma_i(f)}{\sum_{j=1}^M \gamma_j(f)}}. \quad (14)$$

Assuming the noise power is the same for all microphones and substituting $\gamma_i(f)$ by (10) leads to

$$G_{\text{SC}}^{(i)}(f) = \sqrt{\frac{|H_i(f)|^2}{\sum_{j=1}^M |H_j(f)|^2}} = \frac{|H_i(f)|}{\sqrt{\sum_{j=1}^M |H_j(f)|^2}}. \quad (15)$$

Hence, we have

$$G_{\text{SC}}^{(i)}(f) = c_{\text{SC}}(f) |H_i(f)| \quad (16)$$

with

$$c_{\text{SC}}(f) = \frac{1}{\sqrt{\sum_{j=1}^M |H_j(f)|^2}}. \quad (17)$$

We observe that the weight $G_{\text{SC}}^{(i)}(f)$ is proportional to the magnitude of the MRC weights $H_i^*(f)$, because the factor c_{SC} is the same for all M microphone signals. Consequently, coherent addition of the sensor signals weighted with the gain factors $G_{\text{SC}}^{(i)}(f)$ still leads to a combining, where the signal-to-noise ratio at the combiner output is the sum of the input SNR values. However, coherent addition requires an additional phase estimate. Let $\phi_i(f)$ denote the phase of $H_i(f)$ at frequency f . Assuming cophasal addition the estimated speech spectrum is

$$\begin{aligned} \hat{X} &= G_{\text{SC}}^{(1)} e^{-j\phi_1} Y_1 + G_{\text{SC}}^{(2)} e^{-j\phi_2} Y_2 + G_{\text{SC}}^{(3)} e^{-j\phi_3} Y_3 \dots \\ &= \frac{1}{c_{\text{SC}}} X + G_{\text{SC}}^{(1)} e^{-j\phi_1} N_1 + G_{\text{SC}}^{(2)} e^{-j\phi_2} N_2 + \dots \end{aligned} \quad (18)$$

Hence, in the case of stationary signals the term

$$\frac{1}{c_{\text{SC}}(f)} = \sqrt{\sum_{j=1}^M |H_j(f)|^2} \quad (19)$$

can be interpreted as the resulting transfer characteristic of the system. An example is depicted in Figure 3. The upper figure presents the measured transfer characteristics for two microphones in a car environment. Note that the microphones have a high-pass characteristic and attenuate signal components for frequencies below 1 kHz. The lower figure is the curve $1/c_{\text{SC}}(f)$. The spectral combiner equalizes most of the deep dips in the transfer functions from the mouth of the speaker to the microphones while the envelope of the transfer functions is not equalized.

3.3. Magnitude Combining. One challenge in multimicrophone systems with spatially separated microphones is a reliable phase estimation of the different input signals. For a coherent combining of the speech signals, we have to compensate the phase difference between the speech signals at each microphone. Therefore, it is sufficient to estimate the phase differences to a reference microphone, for example, to the first microphone $\Delta_i(f) = \phi_1(f) - \phi_i(f)$, for all $i = 2, \dots, M$. Cophasal addition is then achieved by

$$\hat{X} = G_{\text{SC}}^{(1)} Y_1 + G_{\text{SC}}^{(2)} e^{j\Delta_2} Y_2 + G_{\text{SC}}^{(3)} e^{j\Delta_3} Y_3 \dots \quad (20)$$

But a reliable estimation of the phase differences is only possible in speech active periods and furthermore only for that frequencies where speech is present. Estimating the phase differences

$$e^{j\Delta_i(f)} = \mathbb{E} \left\{ \frac{Y_1(f) Y_i^*(f)}{|Y_1(f)| |Y_i(f)|} \right\} \quad (21)$$

leads to unreliable phase values for time-frequency points without speech. In particular, if $H_i(f) = 0$ for some frequency f , the estimated phase $\Delta_i(f)$ is undefined. A combining using this estimate leads to additional signal distortions. Additionally, noise correlation would distort the phase estimation. A coarse estimate of the phase difference can also be obtained from the time-shift τ_i between the speech components in the microphone signals, for example, using the generalized correlation method [20]. The estimate is then $\Delta_i(f) \approx 2\pi f \tau_i$. Note that a combiner using these phase values would in a certain manner be equivalent to a delay-and-sum beamformer. However, for distributed microphone arrays in reverberant environments this phase compensation leads to a poor estimate of the actual phase differences.

Because of the drawbacks, which come along with the phase estimation methods described above, we propose another scheme. Therefore, we use a two stage combining approach. In the first stage, we use the spectral combining approach as described in Section 3.2 with a simple magnitude combining of the microphone signals. For the magnitude combining the noisy phase of the first microphone signal is adopted to the other microphone signals. This is also obvious in Figure 5, where the phase of the noisy spectrum $e^{j\tilde{\phi}_1(f)}$ is taken for the spectrum at the output of the filter $G_{SC}^{(2)}(f)$, before the signals were combined. This leads to the following incoherent combining of the input signals

$$\begin{aligned}\tilde{X}(f) &= G_{SC}^{(1)}(f)Y_1(f) + G_{SC}^{(2)}(f)|Y_2(f)|e^{j\tilde{\phi}_1(f)} + \dots \\ &\quad + G_{SC}^{(M)}(f)|Y_M(f)|e^{j\tilde{\phi}_1(f)} \\ &= G_{SC}^{(1)}(f)Y_1(f) + G_{SC}^{(2)}(f)|Y_2(f)|e^{j\tilde{\phi}_1(f)} + \dots\end{aligned}\quad (22)$$

The estimated speech spectrum $\tilde{X}(f)$ is equal to

$$\frac{X(f)e^{j\tilde{\phi}_1(f)}}{\tilde{c}_{SC}(f)} \quad (23)$$

plus some weighted noise terms. It follows from the triangle inequality that

$$\frac{1}{\tilde{c}_{SC}(f)} \leq \frac{1}{c_{SC}(f)} = \sqrt{\sum_{j=1}^M |H_j(f)|^2}. \quad (24)$$

Magnitude combining does not therefore guarantee maximum-ratio-combining. Yet the signal $\tilde{X}(f)$ is taken as a reference signal in the second stage where the phase compensation is done. This coherence based signal combining scheme is described in the following section.

4. Coherence-Based Combining

As an example of a coherence based diversity system we first consider the two microphone approach by Martin and Vary [5, 6] as depicted in Figure 4. Martin and Vary

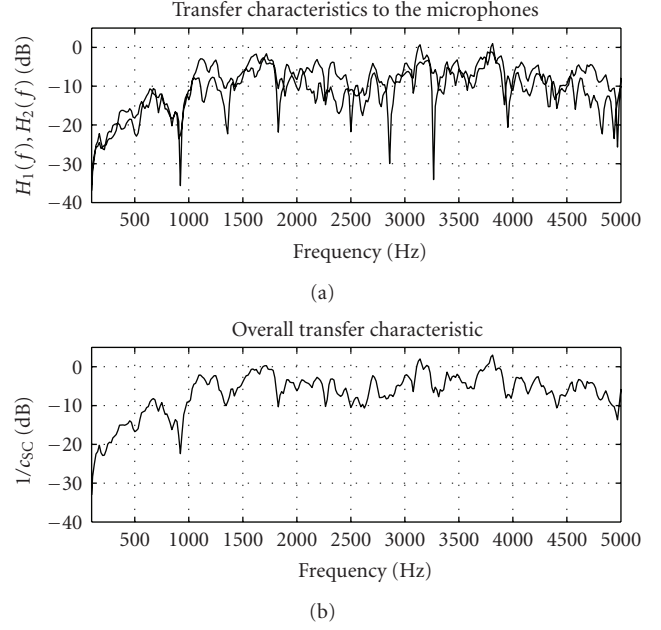


FIGURE 3: Transfer characteristics to the microphones and of the combined signal.

applied the dereverberation principle of Allen et al. [13] to noise reduction. In particular, they proposed an LMS-based time domain algorithm to combine the different microphone signals. This approach provides effective noise suppression for frequencies where the noise components of the microphone signals are uncorrelated.

However, as we have seen in Section 2, for practical microphone distances in the range of 0.4 to 0.8 m the noise signals are correlated for low frequencies. These correlations reduce the noise suppression capabilities of the algorithm and lead to musical noise.

We will show in this section that a combination of the spectral combining with the coherence based approach by Martin and Vary reduces this issues.

4.1. Analysis of the LMS Approach. We present now an analysis of the scheme by Martin and Vary as depicted in Figure 4. The filter $g_i(k)$ is adapted using the LMS algorithm. For stationary signals $x(k)$, $n_1(k)$, and $n_2(k)$, the adaptation converts to filter coefficients $g_i(k)$ and a corresponding filter transfer function

$$G_{LMS}^{(i)}(f) = \frac{\mathbb{E}\{Y_i^*(f)Y_j(f)\}}{\mathbb{E}\{|Y_i(f)|^2\}}, \quad i \neq j \quad (25)$$

that minimizes the expected value

$$\mathbb{E}\left\{\left|Y_i(f)G_{LMS}^{(i)}(f) - Y_j(f)\right|^2\right\}, \quad (26)$$

where $\mathbb{E}\{Y_i^*(f)Y_j(f)\}$ is the cross-power spectrum of the two microphone signals and $\mathbb{E}\{|Y_i(f)|^2\}$ is the power spectrum of the i th microphone signal.

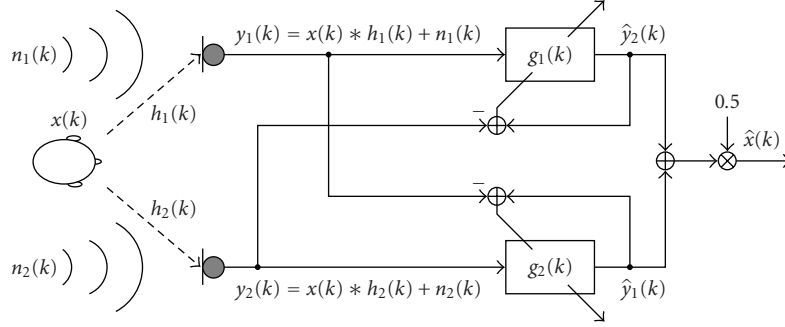


FIGURE 4: Basic system structure of the LMS approach.

Assuming that the speech signal and the noise signals are uncorrelated, (25) can be written as

$$G_{\text{LMS}}^{(i)}(f) = \frac{\mathbb{E}\{|X(f)|^2\}H_i^*(f)H_j(f) + \mathbb{E}\{N_i^*(f)N_j(f)\}}{\mathbb{E}\{|X(f)|^2\}|H_i(f)|^2 + \mathbb{E}\{|N_i(f)|^2\}}. \quad (27)$$

For frequencies where the noise components are uncorrelated, that is, $\mathbb{E}\{N_i^*(f)N_j(f)\} = 0$, this formula is reduced to

$$G_{\text{LMS}}^{(i)}(f) = \frac{\mathbb{E}\{|X(f)|^2\}H_i^*(f)H_j(f)}{\mathbb{E}\{|X(f)|^2\}|H_i(f)|^2 + \mathbb{E}\{|N_i(f)|^2\}}. \quad (28)$$

The filter $G_{\text{LMS}}^{(i)}(f)$ according to (28) results in fact in a minimum mean squared error (MMSE) estimate of the signal $X(f)H_j(f)$ based on the signal $Y_i(f)$. Hence, the weighted output is a combination of the MMSE estimates of the speech components of the two input signals. This explains the good noise reduction properties of the approach by Martin and Vary.

On the other hand, the coherence of the noise depends strongly on the distance between the microphones. For in-car applications, practical distances are in the range of 0.4 to 0.8 m. Therefore, only the noise components for frequencies above 1 kHz can be considered to be uncorrelated [6].

According to formula (27), the noise correlation leads to a bias

$$\frac{\mathbb{E}\{N_i^*(f)N_j(f)\}}{\mathbb{E}\{|Y_i(f)|^2\}} \quad (29)$$

of the filter transfer function. An approach to correct the filter bias by estimating the noise cross-power density was presented in [21]. Another issue with speech enhancement solely based on the LMS approach is that the speech signals at the microphone inputs may only be weakly correlated for some frequencies as shown in Section 2. Consequently, these frequency components will be attenuated in the output signals.

In the following, we discuss a modified LMS approach, where we first combine the microphone signals to obtain an improved reference signal for the adaptation of the LMS filters.

4.2. Combining MRC and LMS. To ensure suitable weighting and coherent signal addition we combine the diversity technique with the LMS approach to process the signals of the different microphones. It is informative to examine the combined approach under ideal conditions, that is, we assume ideal MRC weighting.

Analog to (13), weighting with the MRC gains factors according to (12) results in the estimate

$$\tilde{X}(f) = X(f) + G_{\text{MRC}}^{(1)}(f)N_1(f) + G_{\text{MRC}}^{(2)}(f)N_2(f) + \dots \quad (30)$$

We now use the estimate $\tilde{X}(f)$ as the reference signal for the LMS algorithm. That is, we adapted a filter for each input signal such that the expected value

$$\mathbb{E}\{|Y_i(f)G_{\text{LMS}}^{(i)}(f) - \tilde{X}(f)|^2\} \quad (31)$$

is minimized. The adaptation results in the filter transfer functions

$$G_{\text{LMS}}^{(i)}(f) = \frac{\mathbb{E}\{Y_i^*(f)\tilde{X}(f)\}}{\mathbb{E}\{|Y_i(f)|^2\}}. \quad (32)$$

Assuming that the speech signal and the noise signals are uncorrelated and substituting $\tilde{X}(f)$ according to (30) leads to

$$G_{\text{LMS}}^{(i)}(f) = \frac{\mathbb{E}\{Y_i^*(f)X(f)\}}{\mathbb{E}\{|Y_i(f)|^2\}} \quad (33)$$

$$+ G_{\text{MRC}}^{(i)}(f) \frac{\mathbb{E}\{|N_i(f)|^2\}}{\mathbb{E}\{|Y_i(f)|^2\}}, \quad (34)$$

$$+ G_{\text{MRC}}^{(j)}(f) \frac{\mathbb{E}\{N_i^*(f)N_j(f)\}}{\mathbb{E}\{|Y_i(f)|^2\}} + \dots \quad (35)$$

The first term

$$\frac{\mathbb{E}\{Y_i^*(f)X(f)\}}{\mathbb{E}\{|Y_i(f)|^2\}} = \frac{H_i(f)^* \mathbb{E}\{|X(f)|^2\}}{|H_i(f)|^2 \mathbb{E}\{|X(f)|^2\} + \mathbb{E}\{|N_i(f)|^2\}} \quad (36)$$

in this sum is the Wiener filter that results in a minimum mean squared error estimate of the signal $X(f)$ based on the signal $Y_i(f)$. The Wiener filter equalizes the microphone signal and minimizes the mean squared error between the filter output and the actual speech signal $X(f)$. Note that the phase of the term in (36) is $-\phi_i$, that is, the filter compensates the phase of the acoustic transfer function $H_i(f)$.

The other terms in the sum can be considered as filter biases where the term in (34) depends on the noise power density of the i th input. The remaining terms depend on the noise cross power and vanish for uncorrelated noise signals. However, noise correlation might distort the phase estimation.

Similarly, when we consider the actual reference signal $\tilde{X}(f)$ according to (22), the filter equation for $G_{\text{LMS}}^{(i)}(f)$ contains the term

$$\frac{H_i(f)^* \mathbb{E}\{|X(f)|^2\} e^{j\phi_i(f)}}{\tilde{c}_{\text{SC}}(f) \left(|H_i(f)|^2 \mathbb{E}\{|X(f)|^2\} + \mathbb{E}\{|N_i(f)|^2\} \right)} \quad (37)$$

with the sought phase $\Delta_i(f) = \phi_1(f) - \phi_i(f)$. If the correlation of the noise terms is sufficiently small we obtain the estimated phase

$$\hat{\Delta}_i(f) = \arg\{G_{\text{LMS}}^{(i)}(f)\}. \quad (38)$$

The LMS algorithm estimates implicitly the phase differences between the reference signal $\tilde{X}(f)$ and the input signals $Y_i(f)$. Hence, the spectra at the outputs of the filters $G_{\text{LMS}}^{(i)}(f)$ are in phase. This enables a cophasal addition of the signals according to (20).

By estimating the noise power and noise cross-power densities we could correct the biases of the LMS filter transfer functions. Similarly, reducing the noisy signal components in (30) diminishes the filter biases. In the following, we will pursue the latter approach.

4.3. Noise Suppression. Maximum-ratio-combining provides an optimum weighting of the M sensor signals. However, it does not necessarily suppress the noisy signal components. We therefore combine the spectral combining with an additional noise suppression filter. Of the numerous proposed noise reduction techniques in literature, we consider only spectral subtraction [4] which supplements the spectral combining quite naturally. The basic idea of spectral subtraction is to subtract an estimate of the noise floor from an estimate of the spectrum of the noisy signal.

Estimating the overall SNR according to (9) the spectral subtraction filter (see i.e., [1, page 239]) for the combined signal $\tilde{X}(f)$ can be written as

$$G_{\text{NS}}(f) = \sqrt{\frac{\gamma(f)}{1 + \gamma(f)}}. \quad (39)$$

Multiplying this filter transfer function with (14) leads to the term

$$\sqrt{\frac{\gamma_i(f)}{\gamma(f)}} \sqrt{\frac{\gamma(f)}{1 + \gamma(f)}} = \sqrt{\frac{\gamma_i(f)}{1 + \gamma(f)}} \quad (40)$$

This formula shows that noise suppression can be introduced by simply adding a constant to the numerator term in (14).

Most, if not all, implementations of spectral subtraction are based on an over-subtraction approach, where an overestimate of the noise power is subtracted from the power spectrum of the input signal (see e.g., [22–25]). Over-subtraction can be included in (40) by using a constant ρ larger than one. This leads to the final gain factor

$$G_{\text{SC}}^{(i)}(f) = \sqrt{\frac{\gamma_i(f)}{\rho + \gamma(f)}}. \quad (41)$$

The parameter ρ does hardly affect the gain factors for high signal-to-noise ratios retaining optimum weighting. For low signal-to-noise ratios this term leads to an additional attenuation. The over-subtraction factor is usually a function of the SNR, sometimes it is also chosen differently for different frequency bands [25].

5. Implementation Issues

Real world speech and noise signals are non-stationary processes. For an implementation of the spectral weighting, we have to consider short-time spectra of the microphone signals and estimate the short-time power spectral densities (PSD) of the speech signal and the noise components.

Therefore, the noisy signal $y_i(k)$ is transformed into the frequency domain using a short-time Fourier transform of length L . Each block of L consecutive samples is multiplied with a Hamming window. Subsequent blocks are overlapping by K samples. Let $Y_i(\kappa, \nu)$, $X_i(\kappa, \nu)$, and $N_i(\kappa, \nu)$ denote the corresponding short-time spectra, where κ is the subsampled time index and ν is the frequency bin index.

5.1. System Structure. The processing system for two inputs is depicted in Figure 5. The spectrum $\tilde{X}(\kappa, \nu)$ results from incoherent magnitude combining of the input signals

$$\begin{aligned} \tilde{X}(\kappa, \nu) &= G_{\text{SC}}^{(1)}(\kappa, \nu) Y_1(\kappa, \nu) \\ &+ G_{\text{SC}}^{(2)}(\kappa, \nu) |Y_2(\kappa, \nu)| e^{j\hat{\phi}_1(\kappa, \nu)} + \dots, \end{aligned} \quad (42)$$

where

$$G_{\text{SC}}^{(i)}(\kappa, \nu) = \sqrt{\frac{\gamma_i(\kappa, \nu)}{\rho + \gamma(\kappa, \nu)}}. \quad (43)$$

The power spectral density of speech signals is relatively fast time varying. Therefore, the FLMS algorithm requires a quick update, that is, a large step size. If the step size is sufficiently large the magnitudes of the FLMS filters $G_{\text{LMS}}^{(i)}(\kappa, \nu)$ follow the filters $G_{\text{SC}}^{(i)}(\kappa, \nu)$. Because the spectra at the outputs of the filters $G_{\text{LMS}}^{(i)}(f)$ are in phase, we obtain the estimated speech spectrum as

$$\hat{X}(\kappa, \nu) = G_{\text{LMS}}^{(1)}(\kappa, \nu) Y_1(\kappa, \nu) + G_{\text{LMS}}^{(2)}(\kappa, \nu) Y_2(\kappa, \nu) + \dots \quad (44)$$

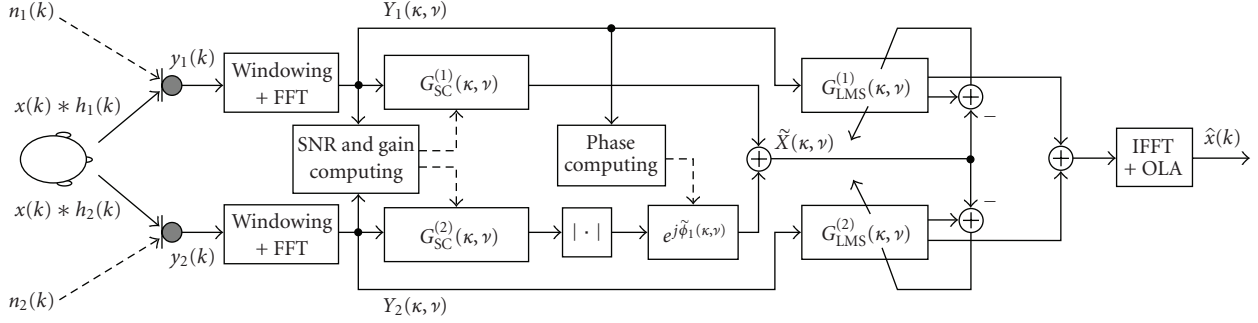


FIGURE 5: Basic system structure of the diversity system with two inputs.

To perform spectral combining we have to estimate the current signal-to-noise ratio based on the noisy microphone input signals. In the next sections, we propose a simple and efficient method to estimate the noise power spectral densities of the microphone inputs.

5.2. PSD Estimation. Commonly the noise PSD is estimated in speech pauses where the pauses are detected using voice activity detection (VAD, see e.g., [24, 26]). VAD-based methods provide good estimates for stationary noise. However, they may suffer from error propagation if subsequent decisions are not independent. Other methods, like the minimum statistics approach introduced by Martin [23, 27], use a continuous estimation that does not explicitly differentiate between speech pauses and speech active segments.

Our estimation method combines the VAD approach with the minimum statistics (MS) method. Minimum statistics is a robust technique to estimate the power spectral density of non-stationary noise by tracing the minimum of the recursively smoothed power spectral density within a time window of 1 to 2 seconds. We use these MS estimates and a simple threshold test to determine voice activity for each time-frequency point.

The proposed method prevents error propagation, because the MS approach is independent of the VAD. During speech pauses the noise PSD estimation can be enhanced compared with an estimate solely based on minimum statistics. A similar time-frequency dependent VAD was presented by Cohen to enhance the noise power spectral density estimation of minimum statistics [28].

For time-frequency points (κ, ν) where the speech signal is inactive, the noise PSD $\mathbb{E}\{|N_i(\kappa, \nu)|^2\}$ can be approximated by recursive smoothing

$$\mathbb{E}\{|N_i(\kappa, \nu)|^2\} \approx \lambda_{Y,i}(\kappa, \nu) \quad (45)$$

with

$$\lambda_{Y,i}(\kappa, \nu) = (1 - \alpha)\lambda_{Y,i}(\kappa - 1, \nu) + \alpha|Y_i(\kappa, \nu)|^2, \quad (46)$$

where $\alpha \in (0, 1)$ is the smoothing parameter.

During speech active periods the PSD can be estimated using the minimum statistics method introduced by Martin

[23, 27]. With this approach, the noise PSD estimate is determined by the minimum value

$$\lambda_{\min,i}(\kappa, \nu) = \min_{l \in [\kappa - W + 1, \kappa]} \{\lambda_{Y,i}(l, \nu)\} \quad (47)$$

within a sliding window of W consecutive values of $\lambda_{Y,i}(\kappa, \nu)$. The noise PSD is then estimated by

$$\mathbb{E}\{|N_i(\kappa, \nu)|^2\} \approx o_{\min} \cdot \lambda_{\min,i}(\kappa, \nu), \quad (48)$$

where o_{\min} is a parameter of the algorithm and should be approximated as

$$o_{\min} = \frac{1}{\mathbb{E}\{\lambda_{\min}\}}. \quad (49)$$

The MS approach provides a rough estimate of the noise power that strongly depends on the smoothing parameter α and the window size of the sliding window (for details cf. [27]). However, this estimate can be obtained regardless of speech being present or not.

The idea of our approach is to approximate the PSD by the MS estimate during speech active periods while the smoothed input power is used for time-frequency points where speech is absent.

$$\begin{aligned} \mathbb{E}\{|N_i(\kappa, \nu)|^2\} &\approx \beta(\kappa, \nu) o_{\min} \cdot \lambda_{\min,i}(\kappa, \nu) \\ &\quad + (1 - \beta(\kappa, \nu)) \lambda_{Y,i}(\kappa, \nu), \end{aligned} \quad (50)$$

where $\beta(\kappa, \nu) \in \{0, 1\}$ is an indicator function for speech activity which will be discussed in more detail in the next section.

The current signal-to-noise ratio is then obtained by

$$\gamma_i(\kappa, \nu) = \frac{\mathbb{E}\{|Y_i(\kappa, \nu)|^2\} - \mathbb{E}\{|N_i(\kappa, \nu)|^2\}}{\mathbb{E}\{|N_i(\kappa, \nu)|^2\}}, \quad (51)$$

assuming that the noise and speech signals are uncorrelated.

5.3. Voice Activity Detection. Human speech contains gaps not only in time but also in frequency domain. It is therefore reasonable to estimate the voice activity in the time-frequency domain in order to obtain a more accurate VAD.

The VAD function $\beta(\kappa, \nu)$ can then be calculated upon the current input noise PSD obtained by minimum statistics.

Our aim is to determine for each time-frequency point (κ, ν) whether the speech signal is active or inactive. We therefore consider the two hypotheses $H_1(\kappa, \nu)$ and $H_0(\kappa, \nu)$ which indicate speech presence or absence at the time-frequency point (κ, ν) , respectively. We assume that the coefficients $X(\kappa, \nu)$ and $N_i(\kappa, \nu)$ of the short-time spectra of both the speech and the noise signal are complex Gaussian random variables. In this case, the current input power, that is, squared magnitude $|Y_i(\kappa, \nu)|^2$, is exponentially distributed with mean (power spectral density)

$$\lambda_{Y_i}(\kappa, \nu) = \mathbb{E}\{|Y(\kappa, \nu)|^2\}. \quad (52)$$

Similarly we define

$$\begin{aligned} \lambda_{X_i}(\kappa, \nu) &= |H_i(\kappa, \nu)|^2 \mathbb{E}\{|X(\kappa, \nu)|^2\}, \\ \lambda_{N_i}(\kappa, \nu) &= \mathbb{E}\{|N_i(\kappa, \nu)|^2\}. \end{aligned} \quad (53)$$

We assume that speech and noise are uncorrelated. Hence, we have

$$\lambda_{Y_i}(\kappa, \nu) = \lambda_{X_i}(\kappa, \nu) + \lambda_{N_i}(\kappa, \nu) \quad (54)$$

during speech active periods and

$$\lambda_{Y_i}(\kappa, \nu) = \lambda_{N_i}(\kappa, \nu) \quad (55)$$

in speech pauses.

In the following, we occasionally omit the dependency on κ and ν in order to keep the notation lucid. The conditional probability density functions of the random variable $\mathcal{Y}_i = |Y_i(\kappa, \nu)|^2$ are [29]

$$f(\mathcal{Y}_i | H_0) = \begin{cases} \frac{1}{\lambda_{N_i}} \exp\left(\frac{-\mathcal{Y}_i}{\lambda_{N_i}}\right), & \mathcal{Y}_i \geq 0, \\ 0, & \mathcal{Y}_i < 0, \end{cases} \quad (56)$$

$$f(\mathcal{Y}_i | H_1) = \begin{cases} \frac{1}{\lambda_{X_i} + \lambda_{N_i}} \exp\left(\frac{-\mathcal{Y}_i}{\lambda_{X_i} + \lambda_{N_i}}\right), & \mathcal{Y}_i \geq 0, \\ 0, & \mathcal{Y}_i < 0. \end{cases} \quad (57)$$

Applying Bayes rule for the conditional speech presence probability

$$p_i(\kappa, \nu) = P(H_1 | \mathcal{Y}_i) \quad (58)$$

we have [29]

$$p_i(\kappa, \nu) = \left\{ 1 + \frac{(\lambda_{X_i} + \lambda_{N_i})q}{\lambda_{N_i}(1-q)} \exp(-u_i) \right\}^{-1}, \quad (59)$$

where $q(\kappa, \nu) = P(H_0(\kappa, \nu))$ is the a priori probability of speech absence and

$$u_i(\kappa, \nu) = \frac{\mathcal{Y}_i \lambda_{X_i}}{\lambda_{N_i}(\lambda_{X_i} + \lambda_{N_i})} = \frac{|Y_i(\kappa, \nu)|^2 \lambda_{X_i}}{\lambda_{N_i}(\lambda_{X_i} + \lambda_{N_i})}. \quad (60)$$

The decision rule for the i th channel is based on the conditional speech presence probability

$$\beta_i(\kappa, \nu) = \begin{cases} 1, & \frac{P(H_1 | \mathcal{Y}_i)}{P(H_0 | \mathcal{Y}_i)} \geq T, \\ 0, & \text{otherwise.} \end{cases} \quad (61)$$

The parameter $T > 0$ enables a tradeoff between the two possible error probabilities of voice activity detection. A value $T > 1$ decreases the probability of a false alarm, that is, $\beta(\kappa, \nu) = 1$ when speech is absent. $T < 1$ reduces the probability of a miss, that is, $\beta(\kappa, \nu) = 0$ in the presence of speech. Note that the generalized likelihood-ratio test

$$\frac{P(H_1 | \mathcal{Y}_i)}{P(H_0 | \mathcal{Y}_i)} = \frac{p_i(\kappa, \nu)}{1 - p_i(\kappa, \nu)} \geq T \quad (62)$$

is according to the Neyman-Pearson-Lemma (see e.g., [30]) an optimal decision rule. That is, for a fixed probability of a false alarm it minimizes the probability of a miss and vice versa. The generalized likelihood-ratio test was previously used by Sohn and Sung to detect speech activity in subbands [29, 31].

The test in inequality (62) is equivalent to

$$p_i(\kappa, \nu)^{-1} = \left\{ 1 + \frac{(\lambda_{X,i} + \lambda_{N,i})q}{\lambda_{N,i}(1-q)} \exp(-u_i) \right\} \leq \frac{1+T}{T}, \quad (63)$$

where we have used (59). Solving for $|Y_i(\kappa, \nu)|^2$ using (60), we obtain a simple threshold test for the i th microphone

$$\beta_i(\kappa, \nu) = \begin{cases} 1, & |Y_i(\kappa, \nu)|^2 \geq \lambda_{N,i}(\kappa, \nu) \Theta_i(\kappa, \nu), \\ 0, & \text{otherwise.} \end{cases} \quad (64)$$

with the threshold

$$\Theta_i(\kappa, \nu) = \left(1 + \frac{\lambda_{N,i}}{\lambda_{X,i}} \right) \log \left(\frac{Tq(1 + (\lambda_{X,i}/\lambda_{N,i}))}{(1-q)} \right). \quad (65)$$

This threshold test is equivalent to the decision rule in (61). With this threshold test, speech is detected if the current input power $|Y_i(\kappa, \nu)|^2$ is greater or equal to the average noise power $\lambda_{N,i}(\kappa, \nu)$ times the threshold $\Theta_i(\kappa, \nu)$. This factor depends on the input signal-to-noise ratio $\lambda_{X,i}/\lambda_{N,i}$ and the a priori probability of speech absence $q(\kappa, \nu)$.

In order to combine the activity estimates for the different input signals, we use the following rule

$$\beta(\kappa, \nu) = \begin{cases} 1, & \text{if } |Y_i(\kappa, \nu)|^2 \geq \lambda_{N,i} \Theta_i \text{ for any } i, \\ 0, & \text{otherwise.} \end{cases} \quad (66)$$

6. Simulation Results

In this section, we present some simulation results for different noise conditions typical in a car. For our simulations we consider the same microphone setup as described in Section 2, that is, we use a two-channel diversity system,

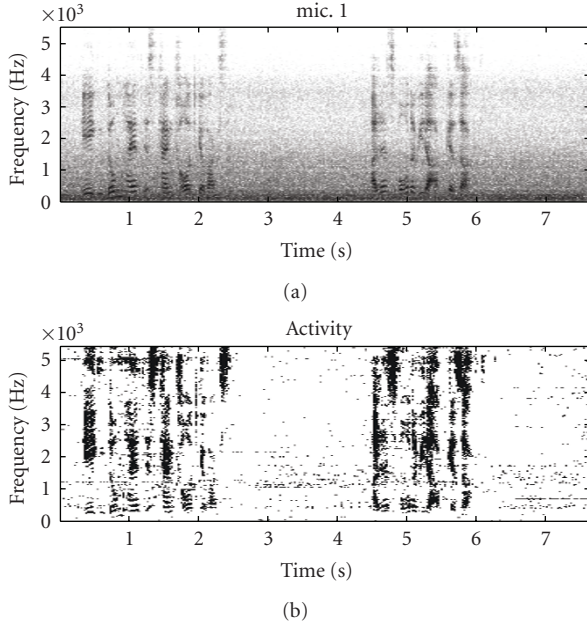


FIGURE 6: Spectrogram of the microphone input (mic. 1 at car speed of 140 km/h, short speaker). The lower figure depicts the results of the voice activity detection (black representing estimated speech activity) with $T = 1.2$ and $q = 0.5$.

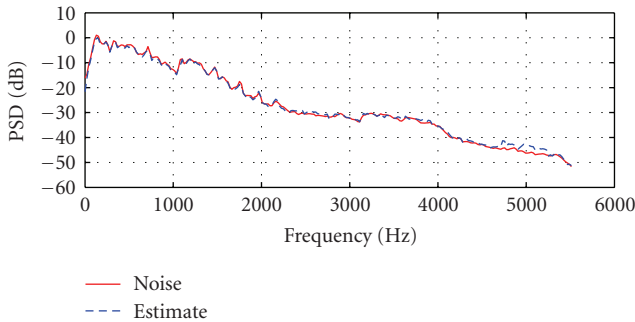


FIGURE 7: Estimated and actual noise PSD for mic. 2 at car speed of 140 km/h.

because this is probably the most interesting case for in-car applications.

With respect to three different background noise situations, we recorded driving noise at 100 km/h and 140 km/h. As third noise situation, we considered the noise which arises from an electric fan (defroster). With an artificial head we recorded speech samples for two different seat positions. From both positions, we recorded two male and two female speech samples, each of a length of 8 seconds. Therefore, we took the German-speaking speech samples from the recommendation P.501 of the International Telecommunication Union (ITU) [32]. Hence the evaluation was done using four different voices with two different speaker sizes, which leads to 8 different speaker configurations. For all recordings, we used a sampling rate of 11025 Hz. Table 1 contains the average SNR values for the considered noise conditions. The first values in each field are with respect to a short speaker

TABLE 1: Average input SNR values [dB] from mic. 1/mic. 2 for typical background noise conditions in a car.

SNR IN	100 km/h	140 km/h	defrost
short speaker	1.2/3.1	-0.7/-0.5	1.7/1.3
tall speaker	1.9/10.8	-0.1/7.2	2.4/9.0

TABLE 2: Log spectral distances with minimum statistics noise PSD estimation and with the proposed noise PSD estimator.

D_{LS} [dB]	100 km/h	140 km/h	defrost
mic. 1	3.93/3.33	2.47/2.07	3.07/1.27
mic. 2	4.6/4.5	3.03/2.33	3.4/1.5

while the second ones are according to a tall person. For all algorithms, we used an FFT length of $L = 512$ and an overlap of 256 samples. For time windowing we apply a Hamming window.

6.1. Estimating the Noise PSD. The spectrogram of one input signal and the result of the voice activity detection are shown in Figure 6 for the worst case scenario (short speaker at car speed of 140 km/h). It can be observed that time-frequency points with speech activity are reliably detected. Because the noise PSD is estimated with minimum statistics also during speech activity, the false alarms in speech pauses do hardly affect the noise PSD estimation.

In Figure 7, we compare the estimated noise PSD with actual PSD for the same scenario. The PSD is well approximated with only minor deviations for high frequencies. To evaluate the noise PSD estimation for several driving situations we calculated as an objective performance measure the log spectral distance (LSD)

$$D_{LS} = \sqrt{\frac{1}{L} \sum_{\nu} \left[10 \log_{10} \frac{\lambda_N(\nu)}{\hat{\lambda}_N(\nu)} \right]^2} \quad (67)$$

between the actual noise power spectrum $\lambda_N(\nu)$ and the estimate $\hat{\lambda}_N(\nu)$. From the definition, it is obvious that the LSD can be interpreted as the mean distance between two PSDs in dB. An extended analysis of different distance measures is presented in [33].

The log spectral distances of the proposed noise PSD estimator are shown in Table 2. The first number in each field is the LSD achieved with the minimum statistics approach while the second number is the value for the proposed scheme. Note that every noise situation was evaluated with four different voices (two male and two female). From these results, we observe that the voice activity detection improves the PSD estimation for all considered driving situations.

6.2. Spectral Combining. Next we consider the spectral combining as discussed in Section 3. Figure 8 presents the output SNR values for a driving situation with a car speed of 100 km/h. For this simulation we used $\rho = 0$, that is, spectral combining without noise suppression. In addition to the output SNR, the curve for ideal maximum-ratio-combining

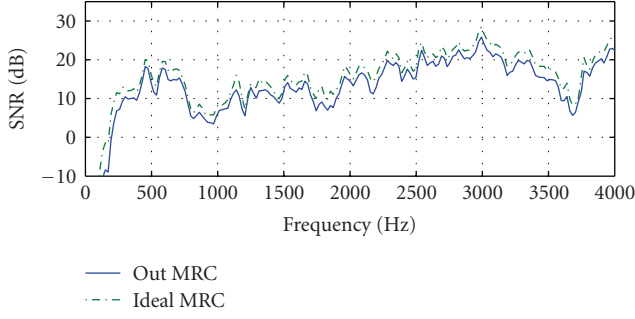


FIGURE 8: Output SNR values for spectral combining without additional noise suppression (car speed of 100 km/h, $\rho = 0$).

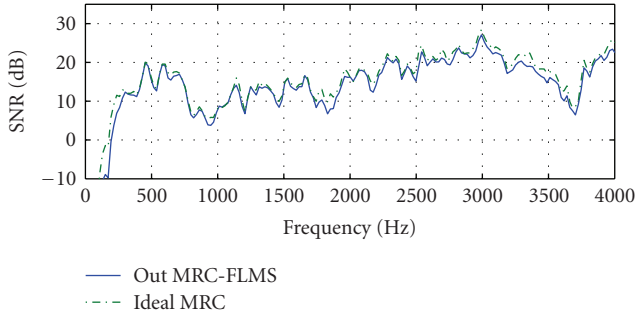


FIGURE 9: Output SNR values for the combined approach without additional noise suppression (car speed of 100 km/h, $\rho = 0$).

is depicted. This curve is simply the sum of the input SNR values for the two microphones which we calculated based on the actual noise and speech signals (cf. Figure 1).

We observe that the output SNR curve closely follows the ideal curve but with a loss of 1–3 dB. This loss is essentially caused by the phase differences of the input signals. With the spectral combining approach only a magnitude combining is possible. Furthermore, the power spectral densities are estimates based on the noisy microphone signals, this leads to an additional loss in the SNR.

6.3. Combining SC and FLMS. The output SNR of the combined approach without additional noise suppression is depicted in Figure 9. It is obvious that the theoretical SNR curve for ideal MRC is closely approximated by the output SNR of the combined system. This is the result of the implicit phase estimation of the FLMS approach which leads to a coherent combining of the speech signals.

Now we consider the combined approach with additional noise suppression ($\rho = 10$). Figure 10 presents the corresponding results for a driving situation with a car speed of 100 km/h. The output SNR curve still follows the ideal MRC curve but now with a gain of up to 5 dB.

In Table 3, we compare the output SNR values of the three considered noise conditions for different combining techniques. The first value is the output SNR for a short speaker while the second number represents the result for the tall speaker. The values marked with FLMS correspond to the coherence based FLMS approach with bias compensation

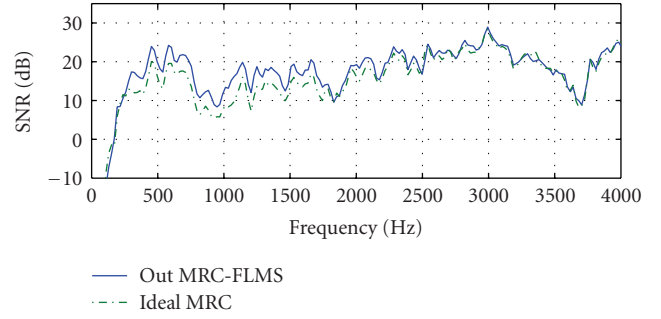


FIGURE 10: Output SNR values for the combined approach with additional noise suppression (car speed of 100 km/h, $\rho = 10$).

TABLE 3: Output SNR values [dB] for different combining techniques—short/tall speaker.

SNR OUT	100 km/h	140 km/h	defrost
FLMS	8.8/13.3	4.4/9.0	7.8/12.3
SC	16.3/20.9	13.3/18.0	14.9/19.9
SC + FLMS	13.5/17.8	10.5/15.0	12.5/16.9
ideal FLMS	12.6/15.2	10.5/13.3	14.5/17.3

TABLE 4: Cosh spectral distances for different combining techniques—short/tall speaker.

D_{CH}	100 km/h	140 km/h	defrost
FLMS	0.9/0.9	0.9/1.0	1.2/1.2
SC	1.3/1.4	1.4/1.5	1.5/1.7
SC + FLMS	1.2/1.1	1.2/1.2	1.4/1.5
ideal FLMS	0.9/0.8	1.1/1.0	1.5/1.4

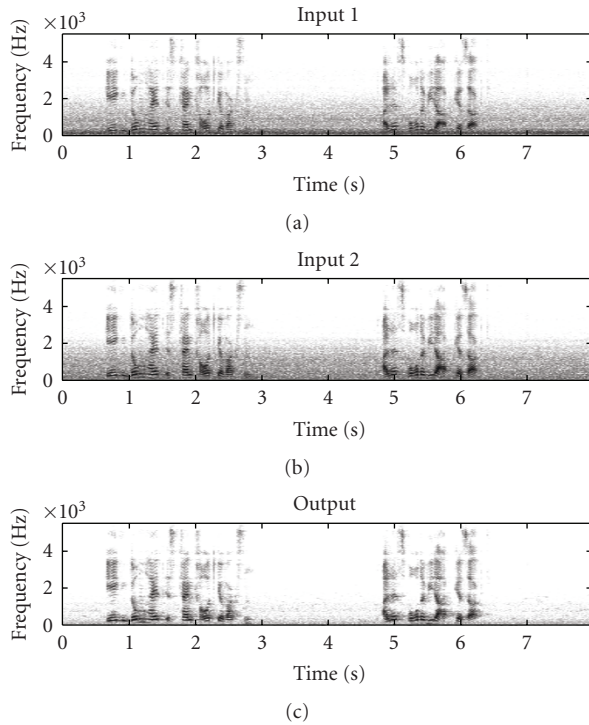
as presented in [21] (see also Section 4.1). The label SC marks results solely based on spectral combining with additional noise suppression as discussed in Sections 3 and 4.3. The results with the combined approach are labeled by SC + FLMS. Finally, the values marked with the label ideal FLMS are a benchmark obtained by using the clean and unreverberant speech signal $x(k)$ as a reference for the FLMS algorithm.

From the results in Table 3, we observe that the spectral combining leads to a significant improvement of the output SNR compared to the coherence based noise reduction. It even outperforms the “ideal” FLMS scheme. However, the spectral combining introduces undesired speech distortions similar to single channel noise reduction. This is also indicated by the results in Table 4. This table presents distance values for the different combining systems. As an objective measure of speech distortion, we calculated the cosh spectral distance (a symmetrical version of the Itakura-Saito distance) between the power spectra of the clean input signal (without reverberation and noise) and the output speech signal (filter coefficients were obtained from noisy data).

The benefit of the combined system is also indicated by the results in Table 5 which presents Mean Opinion Score

TABLE 5: Evaluation of the MOS-Test.

MOS	100 km/h	140 km/h	defrost	average
FLMS	2.58	2.77	2.10	2.49
SC	3.19	3.15	2.96	3.10
SC + FLMS	3.75	3.73	3.88	3.78
ideal FLMS	3.81	3.67	3.94	3.81

FIGURE 11: Spectrograms of the input and output signals with the SC + FLMS approach (car speed of 100 km/h, $\rho = 10$).

(MOS) values for the different algorithms. The MOS test was performed by 24 persons. The test set was taken in a randomized order to avoid statistical dependences on the test order. Obviously, the FLMS approach using spectral combining as reference signal and the “ideal” FLMS filter reference approach are rated as the best noise reduction algorithm, where the values of the combined approach are similar to the results with the reference implementation of the “ideal” FLMS filter solution. From this evaluation, it can also be seen that the FLMS approach with spectral combining outperforms the pure FLMS and the pure spectral combining algorithms in all tested acoustic situations.

The combined approach sounds more natural compared to the pure spectral combining. The SNR and distance values are close to the “ideal” FLMS scheme. The speech is free of musical tones. The lack of musical noise can also be seen in Figure 11, which shows the spectrograms of the enhanced speech and the input signals.

7. Conclusions

In this paper, we have presented a diversity technique that combines the processed signals of several separate microphones. The aim of our approach was noise robustness for in-car hands-free applications, because single channel noise suppression methods are sensitive to the microphone location and in particular to the distance between speaker and microphone.

We have shown theoretically that the proposed signal weighting is equivalent to maximum-ratio-combining. Here we have assumed that the noise power spectral densities are equal for all microphone inputs. This assumption might be unrealistic. However, the simulation results for a two-microphone system demonstrate that a performance close to that of MRC can be achieved with real world noise situations.

Moreover, diversity combining is an effective means to reduce signal distortions due to reverberation and therefore improves the speech intelligibility compared to single channel noise reduction. This improvement can be explained by the fact that spectral combining equalizes frequency dips that occur only in one microphone input (cf. Figure 3).

The spectral combining requires an SNR estimate for each input signal. We have presented a simple noise PSD estimator that reliably approximates the noise power for stationary as well as instationary noise.

Acknowledgments

Research for this paper was supported by the German Federal Ministry of Education and Research (Grant no. 17 N11 08). Last but not the least, the authors would like to thank the reviewers for their constructive comments and suggestions which greatly improve the quality of this paper.

References

- [1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, John Wiley & Sons, New York, NY, USA, 2004.
- [2] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, John Wiley & Sons, New York, NY, USA, 2006.
- [3] E. Hänsler and G. Schmidt, *Speech and Audio Processing in Adverse Environments: Signals and Communication Technology*, Springer, Berlin, Germany, 2008.
- [4] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.
- [5] R. Martin and P. Vary, “A symmetric two microphone speech enhancement system theoretical limits and application in a car environment,” in *Proceedings of the Digital Signal Processing Workshop*, pp. 451–452, Helsingør, Denmark, August 1992.
- [6] R. Martin and P. Vary, “Combined acoustic echo cancellation, dereverberation and noise reduction: a two microphone approach,” *Annales des Télécommunications*, vol. 49, no. 7-8, pp. 429–438, 1994.
- [7] A. A. Azirani, R. L. Bouquin-Jeannès, and G. Faucon, “Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator,” *IEEE Transactions*

- on *Speech and Audio Processing*, vol. 5, no. 5, pp. 484–487, 1997.
- [8] A. Guérin, R. L. Bouquin-Jeannès, and G. Faucon, “A two-sensor noise reduction system: applications for hands-free car kit,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1125–1134, 2003.
 - [9] J. Freudenberger and K. Linhard, “A two-microphone diversity system and its application for hands-free car kits,” in *Proceedings of European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 2329–2332, Lisbon, Portugal, September 2005.
 - [10] T. Gerkmann and R. Martin, “Soft decision combining for dual channel noise reduction,” in *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH—ICSLP '06)*, vol. 5, pp. 2134–2137, Pittsburgh, Pa, USA, September 2006.
 - [11] J. Freudenberger, S. Stenzel, and B. Venditti, “Spectral combining for microphone diversity systems,” in *Proceedings of European Signal Processing Conference (EUSIPCO '09)*, pp. 854–858, Glasgow, UK, July 2009.
 - [12] J. L. Flanagan and R. C. Lummis, “Signal processing to reduce multipath distortion in small rooms,” *Journal of the Acoustical Society of America*, vol. 47, no. 6, pp. 1475–1481, 1970.
 - [13] J. B. Allen, D. A. Berkley, and J. Blauert, “Multimicrophone signal-processing technique to remove room reverberation from speech signals,” *Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, 1977.
 - [14] S. Gannot and M. Moonen, “Subspace methods for multimicrophone speech dereverberation,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1074–1090, 2003.
 - [15] M. Delcroix, T. Hikichi, and M. Miyoshi, “Dereverberation and denoising using multichannel linear prediction,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1791–1801, 2007.
 - [16] I. Ram, E. Habets, Y. Avargel, and I. Cohen, “Multimicrophone speech dereverberation using LIME and least squares filtering,” in *Proceedings of European Signal Processing Conference (EUSIPCO '08)*, Lausanne, Switzerland, August 2008.
 - [17] K. Mukherjee and B.-H. Gwee, “A 32-point FFT based noise reduction algorithm for single channel speech signals,” in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '07)*, pp. 3928–3931, New Orleans, La, USA, May 2007.
 - [18] W. Armbrüster, R. Czarnach, and P. Vary, “Adaptive noise cancellation with reference input,” in *Signal Processing III*, pp. 391–394, Elsevier, 1986.
 - [19] B. Sklar, *Digital Communications: Fundamentals and Applications*, Prentice Hall, Upper Saddle River, NJ, USA, 2001.
 - [20] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.
 - [21] J. Freudenberger, S. Stenzel, and B. Venditti, “An FLMS based two-microphone speech enhancement system for in-car applications,” in *Proceedings of the 15th IEEE Workshop on Statistical Signal Processing (SSP '09)*, pp. 705–708, 2009.
 - [22] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '79)*, pp. 208–211, Washington, DC, USA, April 1979.
 - [23] R. Martin, “Spectral subtraction based on minimum statistics,” in *Proceedings of the European Signal Processing Conference (EUSIPCO '94)*, pp. 1182–1185, Edinburgh, UK, April 1994.
 - [24] H. Puder, “Single channel noise reduction using time-frequency dependent voice activity detection,” in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC '99)*, pp. 68–71, Pocono Manor, Pa, USA, September 1999.
 - [25] A. Juneja, O. Deshmukh, and C. Espy-Wilson, “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 4, pp. 4160–4164, Orlando, Fla, USA, May 2002.
 - [26] J. Ramírez, J. C. Segura, C. Benítez, A. de La Torre, and A. Rubio, “A new voice activity detector using subband order-statistics filters for robust speech recognition,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 1, pp. I849–I852, 2004.
 - [27] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.
 - [28] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.
 - [29] J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaptation,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 1, pp. 365–368, 1998.
 - [30] G. D. Forney Jr., “Exponential error bounds for erasure, list, and decision feedback schemes,” *IEEE Transactions on Information Theory*, vol. 14, no. 2, pp. 206–220, 1968.
 - [31] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.
 - [32] ITU-T, *Test signals for use in telephony, Recommendation ITU-T P.501*, International Telecommunication Union, Geneva, Switzerland, 2007.
 - [33] A. H. Gray Jr. and J. D. Markel, “Distance measures for speech processing,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.